A hands-on introduction to
the Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE)

Christina Tortora, City University of New York (CSI and The Graduate Center) (ctortora@gc.cuny.edu)
Beatrice Santorini, University of Pennsylvania (beatrice@sas.upenn.edu)
Michael Montgomery, University of South Carolina (ullans@yahoo.com)
Frances Blanchette, City University of New York (The Graduate Center) (fblanchette@gc.cuny.edu)

Workshop facilitators:
Christina Tortora, Michael Montgomery, & Frances Blanchette

Workshop objectives:
- To make the Appalachian studies community aware of the various features/components of this in-progress tool, and of what will be available to researchers once the corpus is completed
- To illustrate aspects of the corpus building protocol for those interested in taking on similar projects

1. Introduction

- **Audio-Aligned and Parsed Corpus of Appalachian English (AAPCAppE):** a database that will further research in the various sub-disciplines of Linguistics, and afford novel approaches to the analysis of English dialect data.
- **Ultimate product:** an online, freely accessible, ~1,000,000-word corpus of Appalachian English which will be:
  - syntactically annotated, or “parsed” (searchable by any standard tree query language e.g., *CorpusSearch*, Randall 2009)
  - accompanied by a full set of digitized recordings of the underlying speech signal, in the form of .wav files (text-searchable using Praat (Boersma and Weenink 2011) and ELAN (Wittenburg et al. 2006))
- Based on M. Montgomery’s *Archive of Traditional Appalachian Speech and Culture* (ATASC), an unparsed corpus consisting of the speech and corresponding transcriptions compiled from oral history project recordings housed at various colleges and institutions in the Appalachian region.

Organization of workshop:
- 1:00-1:45 (45 min): Discussion of some basic facts about the corpus, and its building process
- 1:45-1:55 (10 min): Hand-out DVDs; Installation of programs (ELAN and Praat)
- 1:55-2:15 (20 min): EXPLORATION 1 (Praat; *a*-prefixing)
- 2:15-2:30 (15 min): DISCUSSION / QUESTIONS
- 2:30-2:45 (15 min): BREAK
- 2:45-3:05 (20 min): EXPLORATION 2 (ELAN; *uz* vs. *was*)
- 3:05-3:10 (5 min): DISCUSSION / QUESTIONS
- 3:10-3:40 (30 min): EXPLORATION 3 (simple past and participial forms)
- 3:40-3:50 (10 min): DISCUSSION
- 3:50-4:00 (10 min): QUESTIONS AND WRAP UP / RETURN DVDs

2. Some basic facts about building the AAPCAppE

2.1 Background on the ATASC
The AAPCAppE is based on M. Montgomery’s *Archive of Traditional Appalachian Speech and Culture* (ATASC), an unparsed corpus consisting of the speech and corresponding transcriptions compiled from oral history project recordings housed at various colleges and institutions in the Appalachian region.
Some constituent collections of the Archive of Traditional Appalachian Speech and Culture (ATASC):

I. **Joseph Hall Collection.** Interviews with residents of the Great Smoky Mountains in Tennessee and North Carolina (1939); collector: Joseph Hall.

II. **Appalachian Oral History Project (AOHP).** Alice Lloyd College, in Pippa Passes, KY. This history project was conducted from 1971-75 and its materials are housed in the library at Alice Lloyd College, Pippa Passes, Kentucky. The speech is from Central Eastern Kentucky.

III. **Great Smoky Mountains Oral History Project.** Collection of interviews recorded on reels by GSMNP staff and volunteers from 1954 to 1983 with former residents of park lands.

IV. **Dante Oral History Project.** Collection of interviews on cassette tape with residents of Dante, VA (recorded 1997-98), under the auspices of People Incorporated of Abingdon, VA. Recordings are housed at, and curated by, the Archives of Appalachia at East Tennessee State University (ETSU; http://www.etsu.edu/cass/archives/).

V. **Eastern Band Cherokee Collection** at Western Carolina University, in Cullowhee, NC. Interviews done on cassettes from the early 1980s through c2000, by Mountain Heritage Center (http://www.wcu.edu/2389.asp) and colleagues at WCU.

VI. **Appalachian Oral History Project** at Appalachian State University, Boone, NC. Speech of Northwestern North Carolina.

VII. **Augusta Heritage Center Archives.** Davis and Elkins College. German heritage speech, from Eastern West VA; collected in 1990s. Archives hold 1st generation cassette recordings.

VIII. **OTHER PROJECTS PENDING.**

This Workshop is based on the ~50,000 word Joseph Hall sub-corpus. See **Appendix A** for some background on the Joseph Hall recordings.

### 2.2 APPCAppE Protocol

The APPCAppE, which will be ~1-million words when complete, will consist of the following components, all of which will be made available for research:

1. **.wav files of the speech signal**
2. TextGrids — i.e., transcripts which are “time-aligned” with the speech signal (usable in Praat/ELAN)
3. A Part-of-Speech tagged version of the text (searchable with CorpusSearch; Randall 2009)
4. AParsed (or, “syntactically annotated”) version of the text (searchable with CorpusSearch)
5. A complete, basic transcript (for those not interested in the fancy stuff in (1-4))

### 2.2.1 Phonetics/Phonology

#### 2.2.1.1 Digitization

The first part of the corpus creation process requires the digitization of recordings. Here we list a few considerations, questions, and problems which arise:

- **Consideration 1: getting as close to the source material as possible.** Perhaps the wisest move is to digitize off of the original recordings (?) (Copies —and copies of copies — are lower quality, in the case of analog recordings.) NB: Right now our Joseph Hall .wav files are FIFTH “generation”!
  - But not a simple matter: do archivists allow access to originals?
  - And not obviously necessary: experimentation is needed to see if speech signal info is lost in 2nd and 3rd (...and 4th...) generation copies of recordings; perhaps the necessary linguistic information is still present in copies of copies? But how to experiment, i.e., compare digitization of originals to digitization of copies, if one has no access to originals?

- **Consideration 2: How to digitize?** iMic (acts as an external sound card) vs. direct line into computer’s audio-in?
Against iMic or other external sound card: overrides computer’s sound card.

In favor of iMic or other external sound card: using direct audio-in might record noise from computer.

- Consideration 3: What sampling rate to digitize at? “Archive quality” (44100 Hz) seems like a no-brainer, but:
  - Is this overkill for speech (i.e., is there any speech signal information above, say, 16000 Hz (especially in older, taped recordings)?
  - The higher the sampling rate, the bigger the files get; files that are too big are difficult to share, download, load in Praat, archive, etc. (note that some of our files are over 150MB)

2.2.1.2 Aligning transcripts with .wav files in Praat (www.fon.hum.uva.nl/praat/)

We use Praat to provide “time-aligned transcripts” for the .wav files — the TextGrids (will explore in second half of workshop)

Example (1)

There are (at least) two ways one can make a time-aligned transcript, or TextGrid, in Praat:

- Type in the transcript by hand, or
- Use the PPL Forced Aligner (developed by Jiahong Yuan at the U. of Pennsylvania): martinet.sas.upenn.edu/PPLClient/ (this is only an option if you already have a transcript of the recording)

We have thus far used the latter.

The benefits of using the PPL Forced Aligner are:

- The time-aligning is done automatically (so that you do not have to type in the TextGrids by hand). This can save a lot of time.

The drawbacks of using the PPL Forced Aligner are:

- It only makes one TextGrid; so if your recording has more than one speaker, it puts all speakers on the same TextGrid “tier”; this is a problem if you want different tiers for different speakers (which is especially necessary when speakers overlap, and also if you plan on converting the TextGrid files into a text that can be Part-of-Speech-Tagged, for the purposes of building a parsed corpus).
- It places boundaries between each word, as opposed to between each utterance. Putting boundaries between utterances is more useful if one wishes to then transform the TextGrid tiers of more than one speaker into a readable transcript (see e.g. NC State SLAAP project, ncslaap.lib.ncsu.edu/, which provides a script for converting different tiers of speech into a readable Speaker A / Speaker B turn-taking transcript; not possible if there is no indication for utterance start/stop points)

Example (2)

- And finally: A pre-existing transcript can influence your perception! But NB: perception will always be a problem in transcription, no matter what the conditions.

2.2.1.3 ELAN
The TextGrid files can be imported into ELAN (www.lat-mpi.eu/tools/elan/), a program that allows for global text searching throughout the .wav files, allowing the user to call up all tokens of a single item of interest in the corpus (will explore in second half of workshop)

2.2.1.4 General comments about making transcripts / time-aligned TextGrids
Two major problems arise in the transcription of speech:

- **Perceptual problems of Type 1:** The transcriber is a human being, and thus limited (arguably by virtue of genetic endowment) to perceive speech with certain discriminatory mechanisms that prejudice perception. Thus, all humans — linguists and non-linguists alike — have not only a pre-determined “theory of language,” but also a pre-determined “theory of speech” which forces us to conceptualize the acoustic signal in ways that may not line up with physical reality.

  Examples:
  - How many –k’s in the word make in: ...And they **make**, the old timers **made it with with corn** (tape3side2)
  - How many –s’s in the word set in: ...and they’re a dark complected **set** of folks (tape17side2)
  - Is there a –t in once(i), in ...And when a man goes there he doesn’t have the privileges that he **once** had when I was a boy (tape17side2) (this last example illustrates the problem influenced in part by a pre-existing transcription)
If a speaker uses the form *borned* say 3 times (as in: *I was borned in 1863*), it is easy to perceive a fourth token of this verb as *borned*, even if in the fourth token in reality was pronounced *born*.

**NB:** *a*-prefixing (as in *We went a-bear hunting*) is a real hazard in two respects: First, the more a speaker *a*-prefixes with the present participle, the more likely you are to hear an *a*-prefix, even if in reality there are cases where the speaker does not use it. Second, what if the speaker says *uh* a lot? If you hear *I was uh running* ..., how do you decide whether this is *I was a-running*? (tape1 side1, *a*-growing or *uh* growing; *a*-slip aiming or a slip; *a*-*a*-running or *uh*-*a*-running?)

**Perceptual problems of Type 2:** often the linguistic context makes it impossible to perceive the speaker’s intent.
- Examples:
  - *close* vs. *closte*, when followed by *together* or *to*;
  - *I’d dr ank* vs. *I drank*

**Perceptual problems of Type 3:** Variable pronunciation of forms is often gradient.
- Example: Did I hear *was* or *uz*? Did I hear *it* or *hit*?

**Perceptual problems of Type 4:** The recording is simply difficult to hear.
- Did I hear *says* or *said*?
- Did I hear *go to the store* or *go to t'store*? (tape2 side2)

### 2.2.2 Morpho-syntactic

#### 2.2.2.1 Transforming text grid into something that can be read by POS tagger

To make the parsed version of the corpus, we have to go from the TextGrid seen in Example (1) to a transcript which has each and every word and interjection “Part-of-Speech tagged.” This involves taking a file the text of which looks like (3), and transforming it into something that looks like (4):

**Example (3):**
```
<overlap>/CODE
<AppSpeakerA_xmin=43.507>/CODE
Why/WADV
yes/INTJ
of/P
course/N
she/PRO
uh/INTJ
<BREAK>/CODE
</AppSpeakerA_xmax=45.446>/CODE

<Interviewer7_xmin=44.315>/CODE
And/CONJ
how/WADV
<FS>/CODE
What/WD*/WPRO
uh/INTJ
kind/N
of/P
remedies/NS
would/MD
she/PRO
uh/INTJ
give/VB
to/P
the/D
family/N
?/.
</Interviewer7_xmax=48.070>/CODE
</overlap>
```

**Example (4):**
```
<overlap>/CODE
<AppSpeakerA_xmin=43.507>/CODE
Why/WADV
yes/INTJ
of/P
course/N
she/PRO
uh/INTJ
<BREAK>/CODE
</AppSpeakerA_xmax=45.446>/CODE

<Interviewer7_xmin=44.315>/CODE
And/CONJ
how/WADV
<FS>/CODE
What/WD*/WPRO
uh/INTJ
kind/N
of/P
remedies/NS
would/MD
she/PRO
uh/INTJ
give/VB
to/P
the/D
family/N
?/.
</Interviewer7_xmax=48.070>/CODE
</overlap>
```
This requires several steps, some of which are straightforward, and some of which are not.

**Straightforward step:** make a “script” that transforms Example (3) into “tokenable” text (a list of words), as in the following:

(5) Why
   yes
   of
   course
   she
   uh

**Straightforward step:** use POS-tagger that already exists (UPenn), which automatically tags each word with a POS tag.

**Very difficult step:** before feeding (5) into the POS-tagger, must “interleave” speech from different tiers of the Praat TextGrid into a coherent, SpeakerA/SpeakerB turn-taking conversation. *Why do this?*

- [i] we want to transform the TextGrids back into readable transcripts;
- [ii] the Part of Speech tagger works well only if you feed it “coherent” phrases/sentences (i.e., not structures which are interrupted by someone else’s structures)

Regarding [ii]:

Example (6a)
2.2.2 POS tagging/correction

Once the POS-tagger has done its job (tagged each word), we must comb through the result and correct any mistakes.

Many of the mistakes are difficult to predict, and are made by the tagger because it struggles with speech, which is disfluent (e.g., riddled with false starts, repetitions, breaks, etc.). That is, given that the tagger uses statistics to guess the part of speech of a word X (based in part on the fact that words of a particular type precede and follow it), the tagger does worse when the words that precede and follow X are unpredictable.

However, some mistakes are easy to look for and correct.

- **Examples:**
  - The tagger routinely tags *they* as PRO (for pronoun). In Appalachian English, however, it is often EX (for existential), as in *They was just enough of us to fill them three benches.*
  - The tagger routinely tags *them* as PRO (for pronoun). In Appalachian English, however, it is often D (for determiner), as in *And him and them dogs kill that bear.*

Some tags are difficult to know how to correct, as English grammar itself contains forms whose analysis is not clear.

- **Examples:**
  - *I up and told him.* (What tag to give *up*? Present tense verb form? Particle?)
  - *I got two sisters.* (What tag to give *got*? Not a past tense, functionally)
  - *Didn’t he used to do that?* (What tag to give *used*? Not a past tense, functionally, but tagged as a past tense verb in *He used to do that*).
  - *Y’all have a very hard time policing...?* (What tag to give *have*? “bare infinitive” if missing *do*, “simple present” if this is formally a declarative)
  - *I been there...* (What tag to give *been*? as simple past or past participle? **Will return to this**
This “problem” is nevertheless very interesting, as being forced to provide each word with a tag raises important theoretical questions. For example, what is *used to* (*usta*)?

Not a real modal like *will / would /etc.*, which invert in interrogatives:
(7) *Usta he do that?*
    (cf. *Will he do that?*)

Not a modal like *need*, which does not invert, but which is inflected:
(8) I used to / *usta...*  
    He used to / *usta / *ustas...*  
    (cf. *I need vs. He needs / Does he need?*)

Perhaps like invariable/habitual *be* in AAE (Green 2002):
(9) a. Dee *be* waiting for the bus.  
    b. Dee *don’t be* waiting for the bus.

*Used to / *usta*:
(10) a. Dee *used to ( / usta*) wait for the bus.  
    b. Dee *didn’t used to ( / usta*) wait for the bus.

2.2.2.3 Parsing
In order for the corpus to count as “Parsed,” it must be syntactically annotated. The following is an example of the sentence *He sang and danced the polka* (from Santorini’s *Annotation Manual*, www.ling.upenn.edu/~beatrice/annotation/).

(11) *He sang and danced the polka.*

( (IP-MAT (NP-SBJ (PRO He))  
    (VBD sang)  
    ( . . )))

( (IP-MAT (CONJ and)  
    (NP-SBJ *con*)  
    (VBD danced)  
    (NP-OBJ (D the) (N polka))  
    ( . . )))

How does the POS-tagged text (as in example (4) above) come to be parsed? A parser takes the POS-tagged file and automatically assigns syntactic structure.

AAPCAppE: this process is currently in progress: We’re trying to train the parser to handle speech! (biggest challenge: disfluencies such as *false starts, breaks, repetitions*)
3. Hands-on exploration of the AAPCAppE

Three types of hands-on exploration:

(A) Exploration in Praat
(B) Exploration in ELAN
(C) Exploration into Part-of-Speech tagged portion of corpus (POS-tag query demo-site)

Our explorations here will allow us to discuss the kinds of issues/problems/questions which arise in the building of a corpus like the AAPCAppE.

3.1 Phonetics/Phonology in Praat

Here we chose an exploration of a-prefixing, which speaks to the problem of creating completely “truthful” transcripts, i.e., the difficulty of making transcripts which are truly faithful to the speaker’s intent (under Perceptual problems of Type 1, above)

(12) Well back when I were just a young man a-growing up, why we didn’t have any advantage much of schools.

(Ex. from AAPCAppE, found doing a POS-tag query; will turn to this later)

Recall a-prefixing discussion from earlier (for literature on a-prefixing, see e.g. Wolfram 1976, 1980; Wolfram & Christian 1976; Montgomery 2009)

- NB: a-prefixing (as in We went a-bear hunting) is a real hazard in two respects: First, the more a speaker a-prefixes with the present participle, the more likely you are to hear an a-prefix, even if in reality there are cases where the speaker does not use it. Second, what if the speaker says uh a lot? If you hear I was uh running ..., how do you decide whether this is I was a-running ? (tape1side1, a-growing or uh growing; a-(slip)aiming or a slip; a- a-running or uh a-running?)

EXPLORATION: See instructions for opening tape1side1 and tape16side2 in Praat.

Do the following: in each file, search for the following three items:

[i] a= (this is our transcription of an a-prefix)
[ii] a- (this is our transcription of a false start, where the false start is a mid-central vowel)
[iii] uh (this is our transcription of the hesitation uh)

Other possible items to look at:

- How many –s’s in the word set in: ...and they’re a dark complected set of folks (tape17side2)
- Is there a –t in once(t), in ...And when a man goes there he doesn’t have the privileges that he once had when I was a boy (tape17side2) (this last example illustrates the problem influenced in part by a pre-existing transcription)

The above exploration is something one can do easily in ELAN (caveat: we are still working on the bug of ELAN not being able to read the symbol ‘=’). The reason we did this now with Praat is to illustrate Praat’s capability of exhibiting a spectrogram.
3.2 Phonetics/Phonology in \textit{ELAN}

- Advantage of \textit{ELAN} over \textit{Praat}: can do global search throughout a series of files simultaneously
- Disadvantage of \textit{ELAN}: does not give spectrogram

We chose an inquiry which

(a) exploits the global searching capability of \textit{ELAN} (in contrast with \textit{Praat}), and
(b) illustrates again the problem of creating a faithful transcript (and/or the need for phoneticians, who can do better than we can, to analyze such forms!)

Consider \textit{\textipa{az}}, transcribed as =\textit{uz}, as a possible variant of \textit{\textipa{waz} was}:

\begin{enumerate}
\item[(13)]
\begin{enumerate}
  \item ...but he =\textit{uz} scared of the bear.
  \item ...but =\textit{uz} uh steep enough to run good.
\end{enumerate}
\end{enumerate}

\textbf{VS}:

\begin{enumerate}
\item ...And he \textit{was}, he wasn’t just a-running a little.
\item ...And he \textit{was} scared of the bear.
\end{enumerate}

(Exx. from AAPCAppE, \textit{found doing a POS-tag query}; will return to this later)

\textbf{Question}: Why transcribe \textit{\textipa{az}} as =\textit{uz}, as opposed to \textit{\textipa{waz} was}?

\textbf{Answer}: The form \textit{\textipa{az}} is not present in all varieties of English, and as such is worthy of study, in our view. Perhaps there are interesting morpho-phonological or morpho-syntactic constraints on its distribution? Transcribing as many cases as we are certain of (as possible) seems to be a step in the right direction.

However: recall \textbf{Perceptual problems of Type 3}: Variable pronunciation of forms is often gradient.

- Example: Did I hear \textit{was} or \textit{uz}?

\textbf{EXPLORATION}: Search for \textit{uz} in \textit{ELAN}; see instructions for searching through files in \textit{ELAN}.

Other possible items to look at:

- existential \textit{they} vs. existential \textit{there} (are all Appalachian speakers truly always \textit{r-ful}?)
- \textit{hollow}, \textit{fellow}, \textit{panther}, \textit{whip}, once\textit{(t)}, cliff/\textit{clift}, close/\textit{closte}, born/\textit{borned}, throw/\textit{thowed}, says/\textit{said}

3.3 Morpho(-syntax) in \textit{POS-tag query demo-site}

\textbf{Go to}: http://csearch2.ling.upenn.edu/APPALACHIAN/querypos.shtml

Here we chose an exploration of “simple past” (tag: \texttt{VBD}) and “past participial” (tag: \texttt{VBN}) verb forms, which underscores the problems which arise in \textit{POS} tagging — but also, the interesting theoretical questions which arise in tagged/parsed corpus creation.

Specifically, recall from section 2.2.2.2 that \textit{some tags are difficult to know how to correct, as English grammar itself contains forms whose analysis is not clear}.

- Example:
  - \textit{I been there...} (What tag to give \textit{been}? as simple past or past participle?)
Many varieties of English exhibit variability in verb forms, both in simple past contexts, and in present/past perfect contexts (for discussion, see e.g. Wolfram & Fasold 1974; Taylor 1994; Montgomery & Hall 2004)

(14) Forms which are formally identical to both the **bare infinitive** and the **participle** of the Standard:
    *run, come*

(15) Forms which are strictly formally identical to the **bare infinitive** of the Standard:
    *give, kill*

(16) Forms which are strictly formally identical to the **participle** of the Standard:
    *seen, done, been, sung*

(17) Forms which are strictly formally identical to the **simple past** of the Standard:
    *went, saw, took*

(18) Innovative regular forms:
    *seed, knowed, threwed, heared*

Etc.

Many speakers use these forms variably (i.e., there is intra-speaker variability); the Joseph Hall Collection has one speaker who uses *saw, seen, and seed* in a simple past context, all in the same discourse (tape19side1). See Taylor (1994) for the relevance of this fact to morphological theory.

**Tagging:**
We have found it difficult to find a way to tag these forms without being influenced by some theory of what these forms function as, in context (creating a chicken-and-egg problem: we want to tag these forms, so we can study them in order to develop a theory of how a speaker uses them, but we have to have a theory of how the speaker uses them in order to tag...).

There are at least two possible ways to go:

(a) tag the form according to how it is used in the standard variety; e.g., tag *seen* in *I seen him* as a VBN
   
   COMMENT: this makes no sense (for starters: this would lead to the logical conclusion that *walked* in *I walked there* should also be tagged as VBN...)

(b) tag the form according to how the speaker seems to be using it; e.g., tag *seen* in *I seen him* as a VBD, if, for example, there is no overt *have* preceding it.

COMMENT: this seems more sensible. However, the process is not as straightforward as it seems....

*Side note: AppE does exhibit instances of what does seem to be have-deletion:*

(19) *a. I was done supposed to __ been there. (M&H:l)*

    *b. You ought to __ seen us all a-jumping and running (M&H:l)*

    *c. He must’ve died in the forties, must __ been forties whenever he died.

    *d. You wouldn’t __ ever thought about kids a-comin’ out of them hollers and hills.*
Some examples seem clear:

(20) ...and they went down and killed hit, Jess Cable killed it. Come/VBD back, and they =uz just ready to catch another =un, I think. (tape19side1)
   
   ▪ tagged come as VBD, given the context (...but, actually: what about historical present?)

(21) a. He’d went/VBN down the mountain. (tape5side2)
    b. I said that shot’s went/VBN off a- without out any <FS> giving any notice. (tape6side2)
   
   ▪ tagged went as VBN, based on the ’d of He’d and ’s of shot’s

But then what to do in the following cases?

(22) and I says he’s trailed on up on Killpecker and went/VBN to barking. (tape19side1)
   
   ▪ tagged went as VBN, based on the ’s of he’s; TWO PROBLEMS: (a) what is being coordinated here? (b) if we allow for deleted have (see below), this calls into question all cases of went

(23) AppInterviewer: Well, go ahead and tell us about your rating you got there at {inaudible} Creek.
    AppSpeaker: Well I been there about a couple of months, and got the rating thirty six. Next month or so I got <BREAK>
   
   ▪ tagged been as BED, based on the lack of have form, and the surrounding context; PROBLEM: what if there is a deleted had here? (Well, I HAD been there about a couple of months...). In this case, it would be BEN

(24) LATER ON, THE SAME SPEAKERS (tape4side2):
    AppInterviewer: How long you been/BEN a=living on the tow strings, Mister Crisp?
    AppSpeakerA: I’ve been/BEN living on tow string a couple of year.
   
   ▪ tagged first been as BEN, based on the surrounding context; PROBLEM: this suggests a deleted have; so where else should we be hypothesizing a deleted have?

(25) ...Me and a party were on Big Creek fishing, been/BEN out camping out in the in the woods. (tape14side2)
   
   ▪ tagged been as BEN, but, are there strong reasons not to tag as BED?

EXPLORATION:

(a) Go to: http://esearch2.ling.upenn.edu/APPALACHIAN/querypos.shtml

(b) Search for run, come, give, went, seed, saw, seen twice for each form: once as a VBD, once as a VBN; explore our decisions, take count of each form at bottom of page

(c) Search for done, been twice for each form: once as DOD (or BED), once as DON (or BEN); explore our decisions, take count of each form at bottom of page
Acknowledgments:

*We would like to thank Claire Bowern, Marcel den Dikken, Dan Kaufman, Mike Kress, Tony Kroch, Mark Lewental, Ricardo Otheguy, Paul Reed, Edward Snajdr, Doug Whalen, Tiffany Williams, and Jiahong Yuan. The research done thus far has been supported by NSF Grants #BCS-0617197 (Tortora) and #BCS-0616573 (Den Dikken), by an NSF REG Award made to Blanchette on NSF Grant #BCS-0963950 (Snajdr), by The College of Staten Island’s Provost Research Scholarship (2010-11), and by an NEH 2011-12 Fellowship (Tortora). Continued research is to be supported by a National Science Foundation BCS award (2012-16), and by a National Endowment for the Humanities Digital Humanities Start-Up Grant (2012-13) (grant #s TBA).

Appendix A (Michael Montgomery, 2010)
Joseph Hall Recordings of Great Smoky Mountain Speech

(NB: The “transcripts” referred to below are the ATASC transcripts, not the AAPCAppE transcripts)

In 1939 Joseph Sargent Hall (1906-92) spent eight months conducting fieldwork in the Great Smoky Mountains of North Carolina and Tennessee. He had made a preliminary visit to the Smokies in the summer of 1937 under the auspices of the National Park Service, which hired him to begin compiling a record of the traditional culture of residents who were being moved from or given life-time leases on homesteads on which they and their families had lived for sometimes three or four generations. He was expected to focus principally on speech, and as a graduate student at Columbia University studying linguistics at the time, he energetically undertook the challenge. That first summer he used to become acquainted with mountain people, filling several notebooks with observations, but it was not until he returned in 1939 that he began recording them. At the same time he collected a great deal of music. Further information on his work and career can be found in the introduction to Dictionary of Smoky Mountain English, ed. by Michael B. Montgomery and Joseph S. Hall (Knoxville: University of Tennessee Press, 2004) and in “Joseph Sargent Hall: The Man and His Work,” at http://www.cas.sc.edu/engl/dictionary/articles/JosephHall.pdf.

Hall’s recording equipment, transportation, and an assistant, all furnished by the National Park Service and the Civilian Conservation Corps, enabled him to locate speakers throughout the Smokies. He used one CCC camp after another as a base. Local men at each camp identified and often introduced him to “good talkers.” In 1939 he used two recording machines, a Garwick operated by cables hooked to the battery of a pick-up truck, on which he made about 90 aluminum discs, and an Allied that ran on a battery pack, on which he made about 70 acetate discs (these numbers are imprecise because of discrepancies in Hall’s records).

These disk recordings had a short play life and maximum play time of only about five minutes, so to make use of his material for his own research and writing, it was necessary to convert the disks to magnetic tape. At some point in the 1950s he arranged to have this done at the Library of Congress, where he deposited the original disks in exchange for reel copies. These he used to extract anecdotes and other segments that appeared in slightly edited form in two books, Smoky Mountain Folks and Their Lore (1960) and Yarns and Tales from Old Smoky (1978), which appeared from the Cataloochee Press, an imprint which Hall established through a printshop in Asheville, North Carolina. He arranged for a second copy of the twenty reels to be deposited at the Library of Great Smoky Mountains National Park, at the entrance to the Tennessee side of the park near Gatlinburg, so that interpretive staff of the park, scholars, and others could potentially make use of them. These reels were copied onto cassette and subsequently archived in a climate controlled environment. Beyond the excerpts that he published and a few others, Hall did not transcribe any of his recordings.

What truly made Hall’s 1939 recordings accessible was the tireless labor of Mary Ruth Chiles, a retired park employee who took the initiative to transcribe Hall’s recordings and a wealth of others made through an oral history project that interviewed dozens of local residents who had grown up on land taken by the park. Chiles...
transcribed these oral history interviews, which date from 1954 to 1983, as well as other more focused collections of taped interviews such as on the early timber industry in the Smokies and the fiftieth reunion of the CCC. She devised her own system of orthography modified to approximate pronunciation, relying heavily on apostrophes and employing some conventional spellings (e.g. *goin’*, *git*) but others of her own whose phonological quality is sometimes reasonably certain (e.g. *a’right* ‘all right’) but not always (e.g. *th’* ‘the’). To make the recordings more easily searchable, she keyed her transcripts to tape counter numbers every page or two and for most of them prepared subject indexes. She prefaced her transcripts of each tape of Hall’s material with a table of contents, essential because the Library of Congress had collated many disks onto each one.

The transcription below took as its starting point Chiles’ version, but represents a considerable refinement arrived at over a course of many years. With Hall’s permission, Michael Montgomery made copies of the cassettes and transcripts of Hall’s recordings in 1988 and subsequently audited them half a dozen times or more to improve the accuracy of their transcription. Early in this process he decided to use standard orthography wherever possible, except in a small number of cases that usually employ the notation “PRON” to signify an item of noteworthy phonological interest (e.g. *once PRON once* to indicate the presence of a final voiceless consonant, or *recollect PRON recollect*, to indicate that the vowel quality in the initial syllable is [i]). In contrast, items primarily of grammatical rather than phonological interest are normally represented in non-standard spelling (e.g. *borned* ‘born’, *nestes* ‘nests’).

To further ensure as accurate a transcript as possible, Elizabeth Layman (who grew up in Asheville, North Carolina) audited Montgomery’s cassette copies of the recordings in November 2008 using a repeating mechanical transcriber and made additional corrections.

The transcript below is thus based on a fourth-generation recording. Hall’s original disks are in Washington, D.C., and subsequent copies on reel and cassette are in Gatlinburg. These derivative copies of recordings thus may not be as suitable for twenty-first-century sociophonetic analysis of vowel formants as the originals would be. On the other hand, the sound fidelity of Hall’s originals, though not infrequently recorded outdoors, is such that remarkably few (less than one word in a thousand, indicated by “xx”) in the speech of interviewees cannot be ascertained. Beyond inaudible individual words, two problems make the transcript below less than ideal. The recordings are occasionally - but rarely - marred by interruptions, voices repeating themselves, or similar mechanical flaws, and it is impossible to know whether these reflect a mechanical malfunction of Hall’s recording equipment or of the conversion of disk recordings to reels. There are frequent breaks and pauses in the recordings, of varying audibility (indicated by the notation “------” unless a new speaker appears), and without returning to Hall’s 1939 originals, it is impossible to know whether his recorder has been turned off or a disk has run out of space.

The second problem involves transcription of the interviewer’s speech. A soft-spoken man by nature, Hall sought in his interviewing to keep as low and restrained a profile as possible. He can be heard in some interviews (when his turns are indicated by “H.”). As stated earlier, CCC workers directed him to “good talkers,” from whom Hall obtained some idea of what material a speaker could provide before Hall activated his machinery. On many occasions, Hall thus prepped a speaker beforehand, as in asking a man to introduce himself and then recount a exciting bear hunt (see the first speaker below, D. F. Conner) and then Hall would withdraw, so that all one hears is the interviewee and the recording is a performance rather than an interview in any sociolinguistic sense. Hall had few standard questions. He did ask a number of speakers what they thought of the government’s forcing people to leave their land, a subject that might have been a bit more sensitive than Hall realized, coming as it did from a complete stranger, one in fact affiliated with a U.S. Government body. Interestingly, however, answers to questions that created some discomfort has linguistic interest in suggesting something of the politeness codes of mountain people. At other times it almost seems that Hall is very quietly steering the speaker and cannot quite be heard. In these cases, which are of scant if any matter to any auditor today, the notation “------” is also used. One thing that is clear is that on occasion someone besides Hall guided the interview (on several occasions this was Bill Moore). These were men who no doubt were either CCC employees assisting Hall or were speakers whom Hall was aware knew the interviewee well or who could engage in banter with one another.
Hall’s speakers were primarily, but by no means exclusively, older residents. Some of the best “talkers” were middle-aged (e.g. Eugene Sutton, on Tape 16, Side 1, whose anecdotes are featured at the website cited earlier). Occasionally a younger individual is interviewed as well.

It is important not to judge Hall’s interviewing methodology or his selection of speakers on the basis of more modern-day thinking and to realize that many issues that he, a Californian with limited resources and time, had no control over. By comparison to any other student of American English of his day, he collected and put into forms that are now richly usable to researchers a body of material unrivaled for any variety of American English, whether of the Smoky Mountains or anywhere else. His collections, which include notebooks as well as recordings, formed the input to his 1942 Columbia University doctoral dissertation, *The Phonetics of Great Smoky Mountain Speech*. But much beyond his aspiration for a higher degree inspired his work. It was instead his deep love for people and the culture that he found in the Smoky Mountains almost by accident in taking a summer job in 1937 to help pay university expenses. Not only did he return to the Smokies periodically until the 1970s, but he kept in touch with some he befriended until his own passing in 1992. In this latter day his recordings and the transcripts derived from them offer opportunities to historians and many researchers in addition to linguists. Moreover, his work has preserved the voices of ancestors for generations to come far beyond his time and ours.

**Selected Bibliography:**


Icelandic Parsed Historical Corpus, creation in progress, by Joel Wallenberg, Anton Karl Ingason, and Einar Freyr Sigurðsson (http://www.ling.upenn.edu/~joelew/papers/iclt2010treebank.pdf)


SLAAP The Sociolinguistic Archive and Analysis Project (http://ncslaap.lib.ncsu.edu/), North Carolina State University


